ECE 598: Interplay between Control and Machine Learning **Spring 2019**

Lecture 6

Dissipation Inequality for Stochastic Finite-Sum Methods

Lecturer: Bin Hu, Date:02/05/2019

In the last lecture, we have shown that stochastic finite-sum methods can be represented as feedback systems. In Lecture 3, we have presented the dissipation inequality approach as a general analysis tool for feedback systems. In today's lecture, we will discuss how to tailor the dissipation inequality approach for stochastic finite-sum methods.

Suppose G is an LTI system satisfying $\xi_{k+1} - \xi^* = A(\xi_k - \xi^*) + B(w_k - w^*)$. Suppose we know $S = \begin{bmatrix} \xi_k - \xi^* \\ w_k - w^* \end{bmatrix}^\mathsf{T} X \begin{bmatrix} \xi_k - \xi^* \\ w_k - w^* \end{bmatrix} \leq 0$ for any $w = \Delta(C\xi)$.[1] If there exists a positive definite matrix P s.t.

$$\begin{bmatrix} A^\mathsf{T}PA - \rho^2 P & A^\mathsf{T}PB \\ B^\mathsf{T}PA & B^\mathsf{T}PB \end{bmatrix} - X \leq 0, \tag{6.1}$$

then we have $V(\xi_{k+1}) \leq \rho^2 V(\xi_k) + S \leq \rho^2 V(\xi_k)$ where $V(\xi_k) := (\xi_k - \xi^*)^\mathsf{T} P (\xi_k - \xi^*)$. This establishes the linear convergence rate bound $\|\xi_k - \xi^*\| \leq \sqrt{\text{cond}\,(P)}\rho^k \|\xi_0 - \xi^*\|$. We have already discussed how to perform such an analysis for the gradient method. To handle stochastic finite-sum methods, we only need to make some minor modification to the dissipation inequality approach. We first present the high-level ideas.

- **Interconnection of an LTI system G and stochastic Δ:** In this case, one typically will be able to construct some expected supply rate condition $\mathbb{E}S \leq M$. Then the LMI condition (6.1) can still be used to construct a (almost sure) dissipation inequality $V(\xi_{k+1}) \leq \rho^2 V(\xi_k) + S$. How to obtain a convergence bound from such a dissipation inequality? Since the supply rate condition holds in the average sense, we have to take expectation of the dissipation inequality and obtain $\mathbb{E}V(\xi_{k+1}) \leq \rho^2 \mathbb{E}V(\xi_k) + \mathbb{E}S$. Depending on what M is , this expected dissipation inequality can be used to prove various things. For example, when analyzing the stochastic gradient method for smooth strongly-convex f_i, we will figure out that M is just a constant, and the dissipation inequality can be iterated to show $\mathbb{E}V(\xi_k) \leq \rho^2 V(\xi_0) + \frac{M}{1-\rho^2}$. This just states that the stochastic gradient method converges linearly to a small ball whose size is controlled by $\frac{M}{1-\rho^2}$. Notice in this case, the supply rate is not decreasing to 0 and the total internal energy is not going to converge to 0.

- **Interconnection of a jump system G and a deterministic nonlinearity Δ:** As discussed in Lecture 3, we can use the property of Δ to construct some quadratic supply

[1] Here we assume $v_k = C\xi_k$ and hence $w = \Delta(v) = \Delta(C\xi)$.

rate conditions and then analyze the feedback interconnection using the following LMI

$$\sum_{i=1}^{n}\left(p_i\begin{bmatrix}A_i^\mathsf{T}PA_i-\rho^2P & A_i^\mathsf{T}PB_i\\ B_i^\mathsf{T}PA_i & B_i^\mathsf{T}PB_i\end{bmatrix}-X\right)\le 0.$$

Here we assume X is independent of i_k. This type of supply rate conditions arise naturally when the matrix C in G does not depend on i_k and Δ is deterministic. The above LMI can be directly applied to SAGA-like methods.

Now we give a more detailed discussion.

6.1 Incorporating Multiple Supply Rate Conditions

In the dissipation inequality framework, we replace the troublesome perturbation Δ with a quadratic supply rate condition. Relaxing the relation $v=\Delta(w)$ as a quadratic constraint can introduce some conservatism to the analysis. We can reduce the conservatism in the analysis by including multiple supply rate conditions. Suppose we have specified a sequence of symmetric X_j $(j=1,\dots,J)$ such that the following inequalities hold for any (ξ,w) satisfying $w=\Delta(C\xi)$,

$$\begin{bmatrix}\xi_k-\xi^*\\ w_k-w^*\end{bmatrix}^\mathsf{T}X_j\begin{bmatrix}\xi_k-\xi^*\\ w_k-w^*\end{bmatrix}\le 0,\ \forall k$$

Now still consider an LTI model $\xi_{k+1}-\xi^*=A(\xi_k-\xi^*)+B(w_k-w^*)$. We define $S_j(\xi_k,w_k)=\begin{bmatrix}\xi_k-\xi^*\\ w_k-w^*\end{bmatrix}^\mathsf{T}X_j\begin{bmatrix}\xi_k-\xi^*\\ w_k-w^*\end{bmatrix}$. If there exists a positive definite matrix P and non-negative scalers λ_j $(j=1,\dots,J)$ s.t.

$$\begin{bmatrix}A^\mathsf{T}PA-\rho^2P & A^\mathsf{T}PB\\ B^\mathsf{T}PA & B^\mathsf{T}PB\end{bmatrix}\le\sum_{j=1}^{J}\lambda_jX_j,\tag{6.2}$$

then we have $V(\xi_{k+1})\le\rho^2V(\xi_k)+S\le\rho^2V(\xi_k)$ where $V(\xi_k):=(\xi_k-\xi^*)^\mathsf{T}P(\xi_k-\xi^*)$ and $S=\sum_{j=1}^{J}\lambda_jS_j$. The non-negativity of λ_j ensures $S\le 0$. In the LMI (6.2), both P and λ_j are **decision variables**. Therefore, incorporating multiple supply rate conditions just leads to a similar LMI condition with more decision variables.

Why is (6.2) **less conservative than** (6.1)**?** If only one supply rate condition is used (let's say we just use X_1), the resultant LMI condition is just (6.1) with $X=X_1$. In this case, if (6.1) is feasible, then (6.2) is also feasible with $\lambda_1=1$, and $\lambda_j=0$ $(j\ne 1)$ (we just choose the same P). The reverse direction is not true. If (6.2) is feasible, (6.1) with $X=X_1$ may not be feasible. Introducing multiple supply rate conditions helps in many situations. In addition, implementing (6.2) is as easy as implementing (6.1). Therefore, it is almost free to include

extra supply rate conditions if we only care about obtaining numerical rate certifications. Of course, adding more decision variables could cause trouble for analytical rate proofs. Therefore, a more practical way of doing things is to first use numerical implementation to figure out a minimum number of relevant supply rate conditions and then start analytical proofs with those supply rate conditions.

6.2 Dissipation Inequality for Stochastic Gradient

Now we present a detailed analysis for the SG method under the following two assumptions:

1. f is m-strongly convex.
2. f_i is L-smooth and convex for all i.

Under these two assumptions, we can show that SGD satisfies a bound in the following form:

$$\mathbb{E}\|x_k - x^*\|^2 \leq \rho^{2k}\|x_0 - x^*\|^2 + H \tag{6.3}$$

where $\rho^2 = 1 - 2m\alpha + O(\alpha^2)$ and $H = O(\alpha)$. Here ρ^2 quantifies the convergence speed and H quantifies the accuracy. Therefore, for SGD, there is a fundamental trade-off between the convergence speed and the accuracy. If one wants a very accurate solution, one has to decrease α so that H is decreased. However, ρ^2 increases as α decreases and the convergence speed becomes slower.

As mentioned before, the supply rate condition used to prove (6.3) has a form $\mathbb{E}S \leq M$. Recall that the SG method is equivalent to a feedback system $F_u(G, \Delta)$. Here Δ is a stochastic operator mapping v to w as $w_k = \nabla f_{i_k}(v_k)$. In addition, G is governed by an LTI model with $A = I$, $B = -\alpha I$, and $C = I$. We emphasize that for the SG method we have $\xi_{k+1} - \xi^* = A(\xi_k - \xi^*) + Bw_k$ and we do not shift w_k to $(w_k - w^*)$. Again, we perform our analysis in two steps. In Step 1, we construct the supply rates. In Step 2, we solve an LMI to construct the dissipation inequality.

1. Based on $w_k = \nabla f_{i_k}(v_k)$, we can show the following inequalities:

$$\mathbb{E}\begin{bmatrix} v_k - x^* \\ w_k \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} 0 & -LI \\ -LI & I \end{bmatrix} \begin{bmatrix} v_k - x^* \\ w_k \end{bmatrix} \leq \frac{2}{n}\sum_{i=1}^{n}\|\nabla f_i(x^*)\|^2 = M \tag{6.4}$$

$$\mathbb{E}\begin{bmatrix} v_k - x^* \\ w_k \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} 2mI & -I \\ -I & 0 \end{bmatrix} \begin{bmatrix} v_k - x^* \\ w_k \end{bmatrix} \leq 0 \tag{6.5}$$

We skip the proofs here. Now we just set $X_1 = \begin{bmatrix} 0 & -LI \\ -LI & I \end{bmatrix}$ and $X_2 = \begin{bmatrix} 2mI & -I \\ -I & 0 \end{bmatrix}$. Notice that it is the first supply rate that causes the convergence issue for the SG method. Since this supply rate keeps on delivering energy to the system, the internal energy does not decrease to 0.

2. Now we test if there exists $P > 0$ and non-negative scalers (λ_1, λ_2) such that

$$\begin{bmatrix} A^\mathsf{T}PA - \rho^2 P & A^\mathsf{T}PB \\ B^\mathsf{T}PA & B^\mathsf{T}PB \end{bmatrix} - \lambda_1 X_1 - \lambda_2 X_2 \leq 0. \tag{6.6}$$

If so, we have

$$\begin{aligned} &\mathbb{E}(\xi_{k+1} - \xi^*)^\mathsf{T} P(\xi_{k+1} - \xi^*) - \rho^2 \mathbb{E}(\xi_k - \xi^*)^\mathsf{T} P(\xi_k - \xi^*) \\ \leq &\lambda_1 \mathbb{E} \begin{bmatrix} \xi_k - \xi^* \\ w_k \end{bmatrix}^\mathsf{T} X_1 \begin{bmatrix} \xi_k - \xi^* \\ w_k \end{bmatrix} + \lambda_2 \mathbb{E} \begin{bmatrix} \xi_k - \xi^* \\ w_k \end{bmatrix}^\mathsf{T} X_2 \begin{bmatrix} \xi_k - \xi^* \\ w_k \end{bmatrix} \\ \leq &\lambda_1 M \end{aligned}$$

For simplicity, we can choose $P = I$. Recall for SGD we have $A = I$ and $B = -\alpha I$. Hence (6.6) is equivalent to

$$\begin{bmatrix} 1-\rho^2 & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} - \lambda_1 \begin{bmatrix} 0 & -L \\ -L & 1 \end{bmatrix} - \lambda_2 \begin{bmatrix} 2m & -1 \\ -1 & 0 \end{bmatrix} \leq 0 \tag{6.7}$$

Now we set the left side to be a zero matrix. We have $\lambda_1 = \alpha^2$, $\lambda_2 = \alpha - \lambda_1 L$, and $\rho^2 = 1 - 2m\lambda_2 = 1 - 2m\alpha + 2mL\alpha^2$. Now the dissipation inequality leads to

$$\mathbb{E}\|x_{k+1} - x^*\|^2 \leq \rho^2 \mathbb{E}\|x_k - x^*\|^2 + \lambda_1 M$$

Iterating the above bound leads to

$$\begin{aligned} \mathbb{E}\|x_k - x^*\|^2 \leq &\rho^2 \mathbb{E}\|x_{k-1} - x^*\|^2 + \lambda_1 M \\ \leq &\rho^4 \mathbb{E}\|x_{k-1} - x^*\|^2 + (\rho^2 + 1)\lambda_1 M \\ \leq &\rho^{2k} \mathbb{E}\|x_0 - x^*\|^2 + \left(\sum_{t=0}^{\infty} \rho^{2t} \right) \lambda_1 M \\ = &\rho^{2k} \mathbb{E}\|x_0 - x^*\|^2 + \frac{\lambda_1 M}{1 - \rho^2} \end{aligned}$$

From Step 2, we have $\rho^2 = 1 - 2m\alpha + 2mL\alpha^2 = 1 - 2m\alpha + O(\alpha^2)$, and $H = \frac{\lambda_1 M}{1-\rho^2} = O(\alpha)$. This leads to the desired conclusion (6.3).

6.3 Dissipation Inequality for SAGA-like Methods

In the last lecture, we have shown that SAGA and SAG can be rewritten in the feedback form $F_u(G, \Delta)$ where G is some jump system and Δ is a nonlinearity. Specifically, the feedback

interconnection is governed by the following iterative model:

$$\begin{aligned} \xi_{k+1} &= A_{i_k}\xi_k + B_{i_k}w_k \\ v_k &= C\xi_k \\ w_k &= \begin{bmatrix} \nabla f_1(v_k) \\ \nabla f_2(v_k) \\ \vdots \\ \nabla f_n(v_k) \end{bmatrix} \end{aligned} \tag{6.8}$$

We have also shown that the fixed point for SAGA and SAG is given by $\xi^* = \begin{bmatrix}(w^*)^T & (x^*)^T\end{bmatrix}^T$, $w^* = \begin{bmatrix}\nabla f_1(x^*)^T \cdots \nabla f_n(x^*)^T\end{bmatrix}$, and $v^* = x^*$ where x^* satisfies $\frac{1}{n}\sum_{i=1}^n \nabla f_i(x^*) = 0$. For any i_k, the following fixed point condition holds (verify this yourself!)

$$\begin{aligned} \xi^* &= A_{i_k}\xi^* + B_{i_k}w^* \\ v^* &= C\xi^* \\ w^* &= \begin{bmatrix} \nabla f_1(v^*) \\ \nabla f_2(v^*) \\ \vdots \\ \nabla f_n(v^*) \end{bmatrix} \end{aligned} \tag{6.9}$$

To show (6.10) converges to its fixed point, we are actually looking at the following iteration:

$$\begin{aligned} \xi_{k+1} - \xi^* &= A_{i_k}(\xi_k - \xi^*) + B_{i_k}(w_k - w^*) \\ v_k - v^* &= C(\xi_k - \xi^*) \\ w_k - w^* &= \begin{bmatrix} \nabla f_1(v_k) - \nabla f_1(v^*) \\ \nabla f_2(v_k) - \nabla f_2(v^*) \\ \vdots \\ \nabla f_n(v_k) - \nabla f_n(v^*) \end{bmatrix} \end{aligned} \tag{6.10}$$

Again, we can analyze (6.10) by following the two steps in the dissipation inequality framework.

1. First, we try to construct the following supply rate conditions for $j = 1, \ldots, J$.

$$\begin{bmatrix} \xi_k - \xi^* \\ w_k - w^* \end{bmatrix}^{\mathsf{T}} X_j \begin{bmatrix} \xi_k - \xi^* \\ w_k - w^* \end{bmatrix} \leq 0. \tag{6.11}$$

   The supply rate constructions typically require using the matrix C and some properties of Δ. We will cover this in more details in the next few lectures. For now, let's look

at one example. Suppose we know f_1 is L-smooth and m-strongly convex. Hence we know

$$\begin{bmatrix} v_k - v^* \\ \nabla f_1(v_k) - \nabla f_1(v^*) \end{bmatrix}^\mathsf{T} \begin{bmatrix} 2mLI & -(L+m)I \\ -(L+m)I & 2I \end{bmatrix} \begin{bmatrix} v_k - v^* \\ \nabla f_1(v_k) - \nabla f_1(v^*) \end{bmatrix} \leq 0. \quad (6.12)$$

Now notice we have $v_k - v^* = C(\xi_k - \xi^*)$ and $\nabla f_1(v_k) - \nabla f_1(w^*) = (e_1^\mathsf{T} \otimes I)(w_k - w^*)$ where e_1 is a vector whose first entry is 1 and all other entries are 0. Therefore, we have

$$\begin{bmatrix} v_k - v^* \\ \nabla f_1(v_k) - \nabla f_1(v^*) \end{bmatrix} = \begin{bmatrix} C & 0_{p\times(np)} \\ 0 & e_1^\mathsf{T} \otimes I \end{bmatrix} \begin{bmatrix} \xi_k - \xi^* \\ w_k - w^* \end{bmatrix}$$

Substituting the above equation into (6.12) leads to

$$\begin{bmatrix} \xi_k - \xi^* \\ w_k - w^* \end{bmatrix}^\mathsf{T} \begin{bmatrix} C & 0_{p\times(np)} \\ 0 & e_1^\mathsf{T} \otimes I \end{bmatrix}^\mathsf{T} \begin{bmatrix} 2mLI & -(L+m)I \\ -(L+m)I & 2I \end{bmatrix} \begin{bmatrix} C & 0_{p\times(np)} \\ 0 & e_1^\mathsf{T} \otimes I \end{bmatrix} \begin{bmatrix} \xi_k - \xi^* \\ w_k - w^* \end{bmatrix} \leq 0$$

Therefore, we can just choose $X_1 = \begin{bmatrix} C & 0_{p\times(np)} \\ 0 & e_1^\mathsf{T} \otimes I \end{bmatrix}^\mathsf{T} \begin{bmatrix} 2mLI & -(L+m)I \\ -(L+m)I & 2I \end{bmatrix} \begin{bmatrix} C & 0_{p\times(np)} \\ 0 & e_1^\mathsf{T} \otimes I \end{bmatrix}$. Clearly X_1 depends on m, L, and C. You can imagine that properties of f_i and f can all be transformed into quadratic inequalities in the form of (6.11) via similar algebraic manipulations.

2. Now we can perform our LMI-based analysis. If there exists a positive definite matrix P and non-negative scalers λ_j s.t.

$$\sum_{i=1}^{n} \left(p_i \begin{bmatrix} A_i^\mathsf{T} P A_i - \rho^2 P & A_i^\mathsf{T} P B_i \\ B_i^\mathsf{T} P A_i & B_i^\mathsf{T} P B_i \end{bmatrix} \right) \leq \sum_{j=1}^{J} \lambda_j X_j,$$

then we have the expected dissipation inequality $\mathbb{E}V(\xi_{k+1}) \leq \rho^2 \mathbb{E}V(\xi_k) + \mathbb{E}S(\xi_k, w_k)$ where the storage function is defined as $V(\xi_k) = (\xi_k - \xi^*)^\mathsf{T} P (\xi_k - \xi^*)$ and the supply rate S is defined as

$$S(\xi_k, w_k) = \sum_{j=1}^{J} \lambda_j \begin{bmatrix} \xi_k - \xi^* \\ w_k - w^* \end{bmatrix}^\mathsf{T} X_j \begin{bmatrix} \xi_k - \xi^* \\ w_k - w^* \end{bmatrix}.$$

The proof for this part is based on standard Lyapunov arguments you have seen many times. We just left and right multiply both sides of the LMI condition with $\begin{bmatrix} (\xi_k - \xi^*)^\mathsf{T} & (w_k - w^*)^\mathsf{T} \end{bmatrix}$ and $\begin{bmatrix} \xi_k - \xi^* \\ w_k - w^* \end{bmatrix}$. This directly leads to the desired dissipation inequality. The non-negativity of λ_j guarantees $S \leq 0$ and hence we have the linear convergence bound $\mathbb{E}V(\xi_k) \leq \rho^{2k} \mathbb{E}V(\xi_0)$. For given (A_i, B_i, ρ) and X_j, out testing condition is linear in the decision variables P and λ_j, and can be solved as LMIs.

6.4 Further Remarks

Numerically solving LMIs can be done by existing semidefinite program solvers. However, analytically solving the LMIs may require case-by-case constructions of P. The good news is that we can use the numerical solutions of LMIs to guide our constructions of analytical proofs. We will see a few examples in Homework 1.

We can see that once we have the supply rate conditions, the constructions of dissipation inequality can be somehow routinized by solving LMIs. But how to construct supply rates? We have seen a few examples. We will cover the constructions of supply rates in more details in the next few lectures.